Efficient Methods for Incorporating Knowledge into Topic Models

Yi Yang, Doug Downey
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL
yiyang@u.northwestern.edu
ddowney@eecs.northwetsern.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber
@colorado.edu

Ning Li

Outline

- Introduction
- Incorporating Knowledge into LDA
- Experiments
- Conclusion

Introduction

- Topics learned by LDA
  - does not always correlate with human judgments
- incorporate prior knowledge into topic models

Introduction

- Topic modeling with prior knowledge
  - inference is cumbersome for LDA model
  - only work in small-scale scenarios

- Model which can both benefit from rich prior information and scale to large datasets

Incorporating Knowledge into LDA

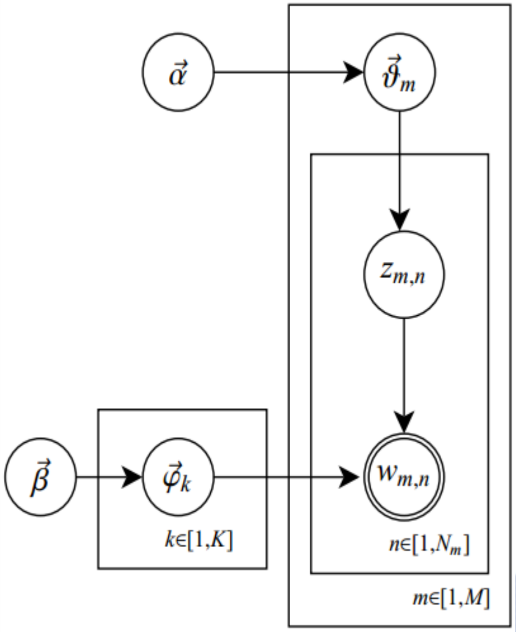
- LDA

θ_d: multinomial distribution over topics for document d

ϕ_z: multinomial distribution over words for topic z

α,β: the hyperparameters for θ and φ(Dirichlet distribution)

$z_{i,j}$: the topic of word j in in the document i from θ_i



Discovering the latent topic assignments z from observed words w requires inferring the the posterior distribution P(z|w)

Incorporating Knowledge into LDA

- collapsed Gibbs sampling

$$P(z = t | \mathbf{z}_{-}, w) \propto (n_{d,t} + \alpha) \frac{n_{w,t} + \beta}{n_t + V\beta} \qquad (1)$$

z_ : the topic assignments of all other tokens w word type
$n_{d,t}$: thenumber of times topic *t* is used in document *d*,
$n_{w,t}$: the number of times word *w* is used in topic *t*,
n_t : the marginal count of the number of tokens assigned to topic t

Incorporating Knowledge into LDA

- collapsed Gibbs sampling

$$\sum_t P(z=t|\mathbf{z}_-, w) = \underbrace{\sum_t \frac{\alpha\beta}{n_t + V\beta}}_{\text{s}} \qquad (2)$$

$$+ \underbrace{\sum_{t, n_{d,t}>0} \frac{n_{d,t}\beta}{n_t + V\beta}}_{\text{r}} + \underbrace{\sum_{t, n_{w,t}>0} \frac{(n_{d,t}+\alpha)n_{w,t}}{n_t + V\beta}}_{\text{q}}.$$

s: the "smoothing only" bucket—constant for all documents
t: the "document only" bucket that is shared by a document's tokens
q: computed specifically for each token,
the sparsity of word-topic count.
the largest mass and few non-zero terms

Incorporating Knowledge into LDA

- Factor Model for Incorporating Prior Knowledge
  - **Existing methods （incorporating prior knowledge） use conventional Gibbs sampling → hinders inference.**
  - **LDA assumes that the hidden topic assignment of a word is independent from other hidden topics. → loses the rich correlation between words.**

Shanghai Jiao Tong University

Incorporating Knowledge into LDA

- Factor Model for Incorporating Prior Knowledge

$$\psi(\mathbf{z}, M) = \prod_{z \in \mathbf{z}} \exp f_m(z, w, d) \qquad (3)$$

M: the set of prior knowledge
z: the current topic assignments
$f_m(z, w, d)$: all hidden topics of word w

If m is knowledge about document d, then $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$ *applies to all topics that are in document d*

f assigns large values to the topics that accord with prior knowledge

Incorporating Knowledge into LDA

- Factor Model for Incorporating Prior Knowledge

$$P(\mathbf{w},\mathbf{z}|\alpha,\beta,M) = P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{z}|\alpha)\psi(\mathbf{z},M) \qquad (4)$$

$$= \int_\theta \int_\phi p(\mathbf{w}|\mathbf{z},\phi)p(\phi|\beta)p(\mathbf{z}|\theta)p(\theta|\alpha)\psi(\mathbf{z},M)d\theta d\phi$$

$$= \psi(\mathbf{z},M) \int_\theta \int_\phi p(\mathbf{w}|\mathbf{z},\phi)p(\phi|\beta)p(\mathbf{z}|\theta)p(\theta|\alpha)d\theta d\phi.$$

$$\hat{P}(\mathbf{z}|\mathbf{w}) = P(\mathbf{z},\mathbf{w}) / \sum_{\mathbf{z}} P(\mathbf{z},\mathbf{w}).$$

Incorporating Knowledge into LDA

- Factor Model for Incorporating Prior Knowledge

· Standard LDA configurations z (set by the hyperparameters α and β) that compromise

  - having few topics per document
  - having few words per topic

Our factor model

adds a further constraint to consider ensembles of

topic assignments z to be compatible with a standard

LDA model and the given prior knowledge

Incorporating Knowledge into LDA

- Factor Model for Incorporating Prior Knowledge

$$
\begin{aligned}
&P(z = t | w, \mathbf{z}_{-}, M) \qquad (5)\\
&= \frac{P(\mathbf{w}, \mathbf{z}_{-}, z = t | \alpha, \beta, M)}{P(\mathbf{w}, \mathbf{z}_{-} | \alpha, \beta, M)}\\
&= \frac{P(\mathbf{w}, \mathbf{z}_{-}, z = t)}{P(\mathbf{w}, \mathbf{z}_{-})} \frac{\psi(\mathbf{z}_{-}, z = t, M)}{\psi(\mathbf{z}_{-}, M)}\\
&\propto \left\{ (n_{d,t} + \alpha) \frac{n_{w,t} + \beta}{n_t + W\beta} \right\} \frac{\psi(\mathbf{z}_{-}, z = t, M)}{\psi(\mathbf{z}_{-}, M)}\\
&\boxed{\propto \left\{ (n_{d,t} + \alpha) \frac{n_{w,t} + \beta}{n_t + W\beta} \right\}} \boxed{\exp f_m(z = t, w, d).}
\end{aligned}
$$

(1) standard LDA
(2) The summation of P(z = t) for sampling.

→

speeding up the sampler is finding a sparse representation

a. word correlation knowledge
b. document-label knowledge

Incorporating Knowledge into LDA

- Word Correlation Prior Knowledge
  - must-link relation : two words tend to be related to the same topics
  - cannot-link relation : two words should not be within the same topic.

Lakers and Celtics → must-link relation

Lakers and bank → cannot-link relation

Incorporating Knowledge into LDA

- Word Correlation Prior Knowledge

$$f_m(z, w, d) = \sum_{u \in M_w^m} \log \max(\lambda, n_{u,z}) + \sum_{v \in M_w^c} \log \frac{1}{\max(\lambda, n_{v,z})}.$$

$\mathrm{M_w}$: a set of prior knowledges **associated with w**
$\mathrm{M_w^m}$: the must-link set of w
$\mathrm{M_w^c}$: the cannot-link set of w
λ : a hyperparameter

Incorporating Knowledge into LDA

- Word Correlation Prior Knowledge

$$
\begin{aligned}
&P(z=t|w,\mathbf{z}_{-},M) \\
&\propto \left\{ \frac{\alpha\beta}{n_t+V\beta} + \frac{n_{d,t}\beta}{n_t+V\beta} + \frac{(n_{d,t}+\alpha)n_{w,t}}{n_t+V\beta} \right\} \\
&\left\{ \prod_{u\in M_w^m} \max(\lambda, n_{u,t}) \prod_{v\in M_w^c} \frac{1}{\max(\lambda, n_{v,t})} \right\}
\end{aligned} \tag{7}
$$

λ : control the “strength” of the prior knowledge term.
If λ is large, the prior knowledge has little impact on the conditional probability of topic assignments

$n_{u,t}$: topic counts for must-link word u
$n_{v,t}$: topic counts for cannot-link word u
→ **All often sparse**

上海交通大學 Shanghai Jiao Tong University

Incorporating Knowledge into LDA

- Other Types of Prior Knowledge
  - **Labeled-LDA (document labels)**

    one-to-one mapping between topics and labels

    restricts topics to be sampled only from the documents label set

$$f_m(z,w,d) = \begin{cases} 1, & \text{if } z \in m_d \\ -\infty, & \text{else} \end{cases}$$

m_d: document d's label set converted to corresponding topic labels

$f_m(z,w,d)$ is sparse

Define $\psi(z, M)$ appropriately so that $f(z, w, d)$ are sparse

Experiments

- Dataset

DATASET	DOCS	TYPE	TOKEN(APPROX)
NIPS	1,500	12,419	1,900,000
NYT-NEWS	3,000,000	102,660	100,000,000
20NG	18,828	21,514	1,946,000

Table 1: Characteristics of benchmark datasets. We use NIPS and NYT for word correlation experiments and 20NG for document label experiments.

Experiments

- Dataset

DATASET	DOCS	TYPE	TOKEN(APPROX)
NIPS	1,500	12,419	1,900,000
NYT-NEWS	3,000,000	102,660	100,000,000
20NG	18,828	21,514	1,946,000

Table 1: Characteristics of benchmark datasets. We use NIPS and NYT for word correlation experiments and 20NG for document label experiments.

Experiments

- Prior Knowledge Generation
  - Word Correlation Prior Knowledge
    - WordNet 3.0 to obtain synsets for word types
    - Existing pretrained word embedding
  - Document Label Prior Knowledge
    - documents in the 20NG dataset are already associated with labels

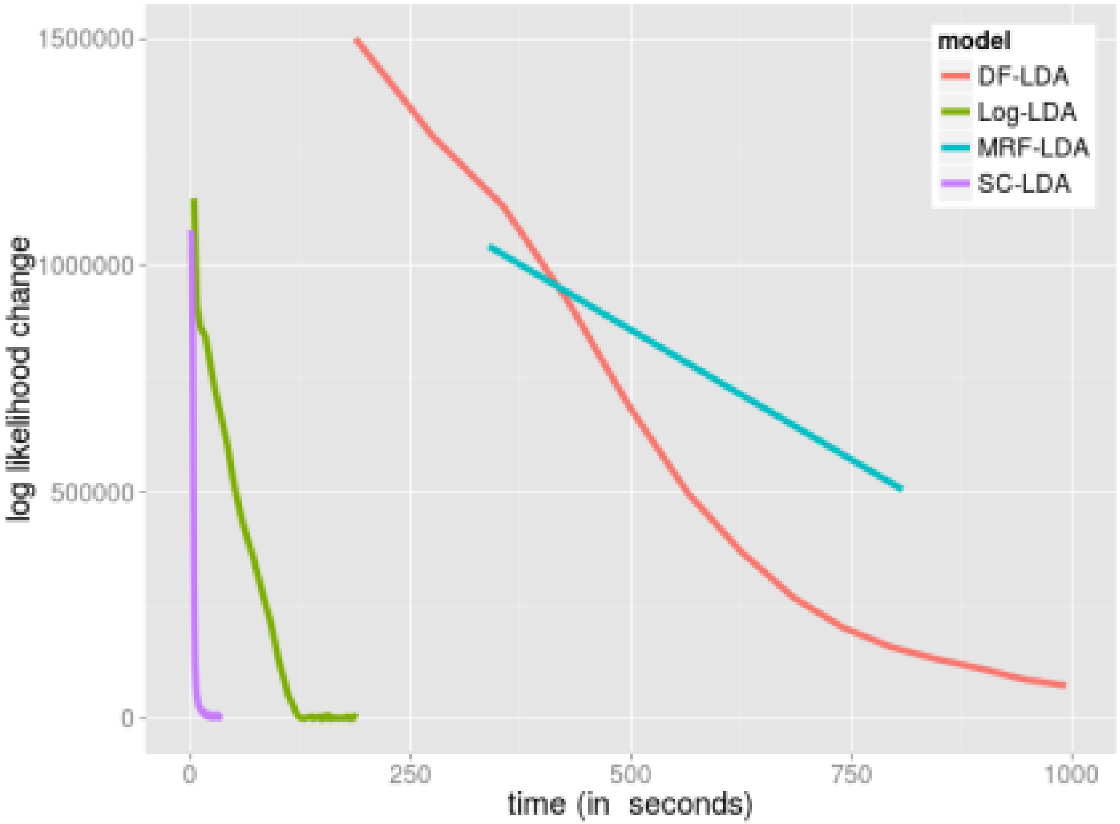
Experiments

- Baselines
  - DF-LDA
    incorporates word must-links and cannot-links using a Dirichlet Forest prior in LDA
  - Logic-LDA
    encodes general domain knowledge as first-order logic and incorporates it in LDA
  - MRF-LDA
    encodes word correlations in LDA as a Markov random field

Experiments

- Convergence

log likelihood change is a good indicator of whether a model has converged or not



Experiments

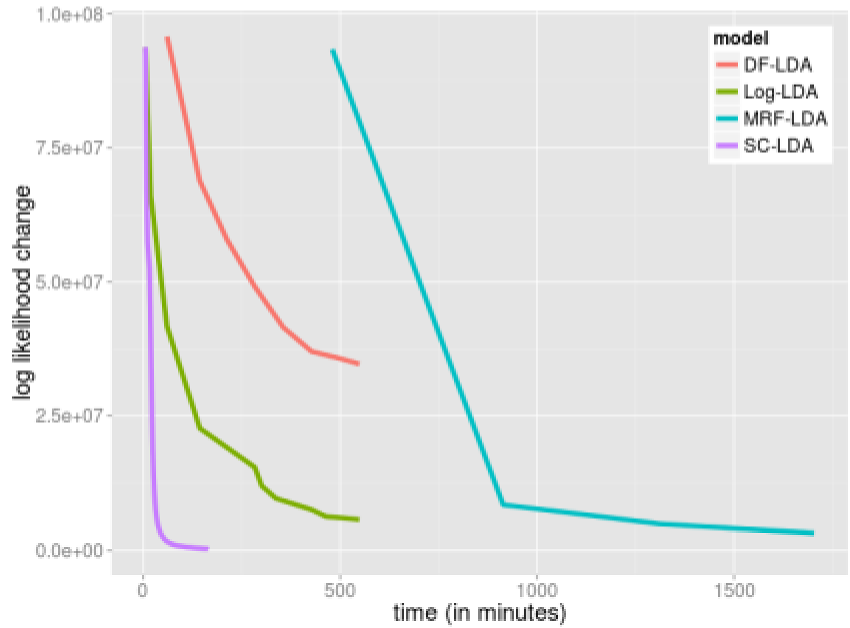


Figure 2: Models' log likelihood convergence on NIPS dataset (above) and NYT-News dataset (below). For NIPS, a 100-topic model with 100 must-links is trained. For NYT-News, a 500-topic model with 100 must-links is trained. SC-LDA reaches likelihood convergence much more rapidly than the other methods.

Experiments

- Convergence

	# Word Correlations			
round	C0	C100	C500	C1000
1st iteration	2.02	2.14	2.30	2.50
50th iteration	0.53	0.56	0.58	0.62
100th iteration	0.48	0.50	0.53	0.56
200th iteration	0.48	0.49	0.52	0.56

Table 2: SC-LDA runtime (in seconds) in the 1st, 50th, 100th, and 200th iteration with different numbers of correlations.

Experiments

- Document Label Prior Knowledge

# Topics				
	T50	T100	T200	T500
Labeled-LDA	0.93	1.89	3.60	8.05
SC-LDA	0.38	0.45	0.51	0.72
# Labeled Documents				
	C500	C1000	C2000	C5000
Labeled-LDA	1.95	1.88	1.75	1.48
SC-LDA	0.51	0.45	0.41	0.31

Table 3: The average running time per iteration over 100 iterations, averaged over 5 seeds, on 20NG dataset. Experiments begin with 100 topics, 1000 labeled documents, and then vary one dimension: number of topics (top), or number of labeled documents (bottom).

Conclusion

- Theory
  - Present a factor graph framework for incorporating prior knowledge into topic models
  - Take advantage of the sparsity to speed up training
- Application(Future direction)
  - Interactive topic modeling
  - Personalized topic modeling

Thank You！